

Information Warfare, Business Intelligence, Text Mining¹

Ing. Alessandro Zanasi

TEMIS Text Mining Solutions

Executive Vice President

Via G.B. Amici, 29-41100 Modena

Alessandro.Zanasi@temis-group.com

Intelligence, dal latino: “Intus Légere”...

ABSTRACT

Gli sviluppi dell'Information Technology hanno rivoluzionato il concetto di guerra, di difesa e di sicurezza. Sempre più spesso si parla infatti di Information Warfare e di Cyber Terrorismo. Ma questi sviluppi hanno anche rivoluzionato il concetto di intelligence e di come fare intelligence.

Grazie infatti ai motori di ricerca e alle nuove tecnologie dei databases, la quantità di informazione disponibile online cresce quotidianamente, mentre la nostra capacità di leggerla ed analizzarla è rimasta praticamente immutata.

Per risolvere il problema che ne deriva una delle tecnologie più innovative al servizio dell'analista di intelligence è, oggi, quella del Text Mining.

Tale tecnologia permette di recuperare l'informazione da centinaia di siti web, agenzie stampa, e-mails, forums, mailing lists, newsgroups e di analizzarla, in pochi secondi, on line.

Grazie al text mining possono essere identificati i temi trattati, le relazioni significative che esistono tra tali temi, possono essere evidenziati i *segnali deboli*, possono essere costruite statistiche sulle informazioni reperite automaticamente, senza alcun intervento umano.

Quest'approccio permette anche di comprendere le opinioni, i gusti, le tendenze del gruppo sociale di cui si sta analizzando la produzione documentale.

PARTE PRIMA: NUOVE FORME DI GUERRA

1. Introduzione

I prossimi conflitti avranno la forma della cosiddetta “information warfare”, ove il “cyber-terrorismo” o i “cyber-crimes” ne rappresenteranno una componente fondamentale. Indicazioni sempre più numerose e frequenti ci indicano che questo tipo di conflitto è già cominciato, sebbene le imprese ed i governi appaiano impreparati ad affrontarlo [Za86].

Governi ed imprese si trovano a dover affrontare individuare le minacce attraverso tecniche di cyber-intelligence.

2. Nuove tecnologie per la guerra

Il mondo sta sperimentando i primi passi di una nuova rivoluzione militare grazie alle cosiddette “nuove tecnologie”. Le tecnologie che la permettono includono:

- le comunicazioni digitali, che permettono ai dati di essere compressi;

¹ Presentato ad Infosecurity – 12 Febbraio 2003.

- un “sistema di posizionamento globale” (GPS), che rende possibile una guida ed una navigazione più precisa;
- sistemi d’arma che possono sfuggire alla rilevazione radar (“stealth”);
- e, naturalmente, l’information technology (IT). In particolare i nuovi strumenti per il recupero e l’analisi automatica dei dati (“data e text mining”).

Tre sono stati gli sviluppi dell’Information Technology che hanno avuto il maggior impatto sugli affari militari.

Il primo riguarda la capacità di raccolta dei dati. Intercettatori su reti telematiche, sensori satellitari o aerei spia possono monitorare virtualmente ogni movimento o discorso o comunicazione in una certa area, raccogliendo grandi volumi di dati [Ma01].

Il secondo riguarda la capacità di analizzare queste moli di dati in maniera automatica. Sistemi avanzati di comando, controllo, comunicazione e computing (C4) danno senso ai dati raccolti dai sensori e mostrati sugli schermi.

E’ qui che si sono avuti i maggiori sviluppi, grazie ai sistemi di data e text mining che permettono di superare il problema dell’eccessiva informazione (information overload), estraendo la conoscenza insita in questi dati senza obbligare l’analista ad avanzare ipotesi sul loro contenuto.

Il terzo consiste nella capacità di mettere in azione tutta questa intelligence-in particolare utilizzando la capacità di entrare in sistemi protetti [St01].

3. Guerra multidimensionale

La nuova guerra sarà multidimensionale, indicando con ciò che non solo le operazioni aeree, marittime e di terra saranno sempre più integrate, ma che queste operazioni riguarderanno anche l’informazione e lo spazio satellitare. “Information warfare” significa anche rendere impotente un nemico penetrando, controllando od anche distruggendo il suo sistema elaborativo di controllo politico, finanziario, telecomunicativo e di gestione di traffico aereo [Gr97].

I governi occidentali temono che un “information strike” possa mettere in ginocchio i loro computers. La US National Security Agency ha assunto centinaia di persone per trattare i problemi dell’information warfare. Ma la decentralizzazione delle moderne reti di computers e la consuetudine di copiare i databases (backup) [Su97] rende improbabile che con un “colpo” solo si possa rendere impotente un’economia occidentale.

E per quanto riguarda la capacità di offesa nemica? Nessun paese potrebbe permettersi ora di costruire un “sistema di sistemi d’arma” capace di competere con quello USA. E neppure potrebbe imboccare la strada di investire pesantemente in armi tradizionali, come in carri armati o in aerei, che sarebbero eccessivamente vulnerabili agli attacchi nemici. Un paese povero, piuttosto, potrebbe scegliere metodi economici, ripugnanti e difficili da contrastare come attentati terroristici, armi chimiche o biologiche [Co97].

E’ per questo che nuove forme di intelligence devono essere pensate per identificare e controllare queste nuove minacce.

4. Guerre senza soldati

Le forze armate di tutti i paesi del mondo dovranno adattarsi alle nuove tecnologie. Sempre meno soldati, marinai ed aviatori combatteranno veramente. E sempre meno gli incaricati della raccolta dei dati per l’intelligence opereranno direttamente in contatto con il nemico.

Saranno sempre di più specialisti in aree come missilistica, informatica, guerra spaziale e, per quanto riguarda l’intelligence, reperimento automatico dell’informazione: data e text mining.

La rivoluzione militare permetterà, in teoria, ad un soldato di saperne sulle operazioni sul campo di battaglia o sulle informazioni strategiche tanto quanto il suo generale.

5. Nuove forme di Terrorismo

Sul finire del XIX secolo, sembrava che nessuno fosse al sicuro da attacchi terroristici. Nel 1894 il presidente francese Sadi Carnot fu ucciso da un anarchico italiano, nel 1897 l'imperatrice austriaca Elizabeth ed il primo ministro spagnolo Canovas furono colpiti a morte dagli anarchici. Nel 1898 l'imperatrice d'Austria Elisabetta, moglie di Francesco Giuseppe, nel 1900 il re Umberto I in Italia e nel 1901 il presidente McKinley negli USA furono a loro volta uccisi. Nel 1909 ad Harbin in Manciuria, il principe giapponese Ito. Nel 1911 il primo ministro russo Stolypin. Nel 1912, di nuovo un Presidente del Consiglio spagnolo, Canalejas.

Visto da questa prospettiva, l'attuale fenomeno del terrorismo non risulta particolarmente preoccupante. Il terrorismo è stato definito come l'utilizzo della violenza con lo scopo di seminare il panico nella società, per indebolirne o sovvertirne gli ordinamenti. Talvolta sfocia nella guerra di guerriglia e talvolta rappresenta un sostituto della guerra tra gli stati.

Il più grande cambiamento negli ultimi decenni per quanto riguarda il terrorismo riguarda il fatto che non è più l'unica strategia a disposizione dei militanti, che sono ora spesso organizzati con un braccio "politico", responsabile dei servizi sociali e della formazione, della gestione degli affari economici e della contestazione delle elezioni, ed un braccio "militare", incaricato delle azioni violente, da cui il braccio politico si dissocia quando queste sono particolarmente gravi e ripugnanti.

L'altro cambiamento riguarda le operazioni terroristiche vere e proprie. Se i dirottamenti aerei sono diventati più rari, le azioni che comportano uccisioni indiscriminate sono aumentate. Inoltre la linea di divisione tra terrorismo urbano ed altre tattiche è diventata meno distinta, come anche quella tra terrorismo politico e criminalità organizzata, almeno per chi non vi è direttamente coinvolto, in special modo per quanto riguarda la Russia e l'America Latina.

Ma il cambiamento maggiore sta proprio nella tecnologia bellica a disposizione di questi nuovi gruppi terroristici. Oltre alle armi nucleari le armi di distruzione di massa includono gli agenti biologici e i composti chimici che attaccano il sistema nervoso, la pelle o il sangue. La possibilità di produrne e di reperirne abbastanza facilmente è la conseguenza della proliferazione di questi ultimi decenni.

Le nuove armi sono alla portata di questi nuovi gruppi, che hanno anche la possibilità di ordinare via posta armi rapidamente disponibili, sia convenzionali che non (la "bomba nucleare dei poveri", come la definì il presidente iraniano Rafsanjani).

Nell'aprile del 1996 un rapporto del dipartimento della difesa USA asseriva che "la maggioranza dei gruppi terroristici non hanno le risorse finanziarie e tecnologiche per acquistare armi nucleari ma potrebbero raccogliere materiale per inquinare radioattivamente, biologicamente e chimicamente ampie regioni". Effettivamente alcuni gruppi sono in grado di ottenere armi di questi tre tipi. In uno degli attacchi più noti di questi ultimi anni, la setta Aum Shinrikyo sparse nella metropolitana di Tokio il gas nervino sarin, uccidendo 10 persone e ferendone 5.000. Tentativi più dilettanteschi in USA ed altrove hanno comportato l'utilizzo della tossina che causa il botulismo, del ricin (due volte), del sarin (due volte), del batterio della peste bubbonica, del batterio del tifo, del vx (altro gas nervino) e, forse, del virus Ebola.

Se i terroristi fino ad oggi hanno utilizzato una sola volta armi chimiche e mai materiale nucleare, in qualche misura le ragioni sono tecniche.

La letteratura scientifica è piena di descrizioni tecniche su come produrre, costruire, immagazzinare e utilizzare armi non convenzionali.

Gli agenti chimici sono molto più facili da produrre o sintetizzare delle armi nucleari ma non sono così facili da controllare e tenere in condizioni stabili, e la loro diffusione dipende largamente dalle condizioni ambientali. I terroristi che organizzarono l'attentato di Tokio scelsero un bersaglio tecnicamente corretto (luogo chiuso con elevata concentrazione di popolazione) ma apparentemente il loro sarin era eccessivamente diluito.

Gli agenti biologici sono di gran lunga più pericolosi: possono uccidere centinaia di migliaia di persone in una sola azione, quando gli agenti chimici ne possono uccidere migliaia.

Relativamente facili da ottenere, sono molto più problematici da stoccare e diffondere ed il rischio che i responsabili della loro diffusione siano contaminati a loro volta è altissimo, nonché la possibilità che questi agenti sopravvivano bene al di fuori dei laboratori ove sono prodotti è ridotta.

La setta Aum Shinrikyo rilasciò il batterio dell'antrace in due occasioni da un edificio a Tokio, senza provocare danni.

6. I nuovi terroristi

Se nel passato il terrorismo era quasi sempre responsabilità di gruppi di militanti di forze politiche ben identificabili, nel futuro il terrorismo sarà appannaggio sempre più di piccoli gruppi o addirittura di individui isolati, sull'esempio di Unabomber, il terrorista americano che, odiando la tecnologia [FC95] e apparentemente operando da solo, per quasi vent'anni inviò pacchi bomba uccidendo o ferendo gravemente molti cittadini USA, o del responsabile dell'attentato agli edifici federali USA di Oklahoma City del 1995. Un individuo infatti può possedere, anche se solo, le competenze tecniche che gli necessitano per rubare, comprare o costruire le armi di cui abbisogna per i suoi obiettivi terroristici. Le ideologie professate da quest'individuo saranno poi probabilmente più aberranti di quelle professate da gruppi maggiori. E questo tipo di terrorista sarà molto più difficile da individuare con i mezzi di intelligence tradizionale, a meno che non incorra in qualche grave errore o sia scoperto per caso [Za87].

7. Cyber-terrorismo

La società è divenuta vulnerabile ad un nuovo genere di terrorismo, in cui il potere distruttivo sia del terrorista singolo che dei mezzi a sua disposizione è divenuto molto maggiore. I terroristi di una volta potevano sì uccidere re e presidenti, ma la loro azione ben poco poteva sulla semplice struttura della società che questi ultimi rappresentavano. Al giorno d'oggi invece la società è molto strutturata e simile ad un organismo complesso la cui esistenza è dipendente dalla sua capacità di produzione, stoccaggio, recupero, analisi e trasmissione dell'informazione. In definitiva è una società più vulnerabile.

La difesa, la polizia, l'attività bancaria, l'attività scientifica ed una larga percentuale delle transazioni dei governi e del settore privato sono on-line. Questo significa che aree vitali della vita delle nazioni sono vulnerabili ad attacchi, frodi e sabotaggi da parte di qualunque hacker sufficientemente esperto e che il sabotaggio organizzato potrebbe impedire ad un paese di funzionare normalmente. Da qui le crescenti speculazioni sull'infoterrorismo e la cyber-guerra [TT93].

PARTE SECONDA: NUOVE FORME DI INTELLIGENCE

8. Le intenzioni politiche e la loro identificazione.

L'ambiente operativo della guerra fredda in cui erano chiare le intenzioni politiche ma non le capacità belliche della controparte si è trasformato in un ambiente in cui le capacità sono chiare (essendo ormai, grazie all'information technology, pubbliche), ma le intenzioni no. E dato che cercare di valutare le capacità tecniche è diventato un esercizio quasi inutile data la rapidità con cui le potenzialità belliche mutano, essere in grado di valutare le intenzioni è diventato sempre più importante.

Dato che proteggere tutte le informazioni è impossibile e che la diffusione dell'informazione aumenta la probabilità che dati importanti, anche se solo in forma parziale, fuoriescano dai circuiti confidenziali dove nascono, l'utilizzo del data e text mining sui dati pubblicamente disponibili ci permette di collegare tra di loro alcuni di essi, permettendoci la ricostruzione dell'informazione strategica.

9. Text Mining: una prospettiva militare

Il text mining permette di trattare i documenti con strumenti di analisi automatica.

Questi strumenti variano considerevolmente tra di loro ma, in generale, riassumono e categorizzano i documenti, identificano la lingua in cui sono scritti, ne estraggono keywords, nomi propri e frasi con più

parole, riportano frequenze di parole e frasi, classificano un documento in funzione della rilevanza rispetto ad uno specifico argomento.

Alcuni incorporano capacità di web-crawling, estraggono dati in varie dimensioni, collegano le informazioni in relazioni spaziali o temporali, scoprono legami o catene di informazioni legate fra di loro, raggruppano documenti in funzione del loro contenuto, effettuano analisi incrociate e permettono l'inclusione di packages statistici.

Attraverso il text mining si possono analizzare volumi immensi di informazioni, sia in tempo reale che in differita e si possono identificare relazioni e strutture che altrimenti sfuggirebbero alla capacità analitica umana.

Le elites che si occupano di sicurezza esprimono idee e pensieri sotto l'influenza di analisi di altissimo livello, briefings militari e diplomatici, gruppi di lavoro interagenti ed incaricati della formulazione della politica nazionale, rapporti di commissioni ed altre fonti dense di informazione.

Queste idee e pensieri contengono perciò tracce del processo articolato attraverso il quale si va costruendo e palesando una decisione strategica.

L'analisi contemporanea di molte fonti relative al medesimo argomento può rivelare strutture e legami che permettono di ricostruire le guidelines attraverso le quali si è pianificata una certa politica ed anche di prevedere le decisioni future.

Il livello migliore su cui rivolgere l'analisi qualitativa del text mining probabilmente non è il vertice del potere, ma gli ambienti a lui vicino.

Molte dichiarazioni di leaders sono palesate in anticipo, anche attraverso osservazioni che sembrano casuali dei rappresentanti del loro staff, e quando il leader comunicherà pubblicamente una politica, quella sarà in verità già stata avviata.

A seconda dell'ambiente, le fonti di informazioni più ricche possono essere all'interno di due o tre cerchi concentrici di persone che si trovano intorno al responsabile di un gruppo.

Questi circoli elitari preparano i briefings decisionali, frequentano gli incontri dei gruppi di lavoro, preparano le bozze che indirizzano i decision makers e strutturano la politica da seguire.

Il Ten.Col.Bill Flynt dell'esercito USA in un suo articolo apparso su *Military Review* del 7 Agosto 2000 [F100] esamina l'applicazione del processo di text mining al libro *Unrestricted Warfare*, scritto da due colonnelli dell'esercito cinese, Qio Ling e Wang Xiangsui.

Sebbene i risultati migliori del text mining si abbiano quando le fonti informative sono numerose, già in questo caso Flynt ne dimostra l'utilità per riconoscere, attraverso l'analisi automatica delle idee espresse dagli autori, le diverse minacce emergenti nelle nuove forme di guerra, i mezzi utilizzati per realizzarle, gli obiettivi ed i fini strategici che si pensa di raggiungere. Come anche mette in evidenza le relazioni che gli autori individuano come essenziali per raggiungere certi obiettivi strategici.

Il Ten.Col. Flynt sottolinea l'importanza del text mining per individuare le prospettive strategiche di un certo soggetto sia che si tratti dell'esercito cinese sia che si tratti, secondo le sue testuali parole, di un terrorista come Unabomber o dell'autore di una pagina web.

10. Informazione elaborata

Come abbiamo più volte detto, l'informazione è un'arma. In particolare lo è oggi e lo è in particolare l'informazione che si trova sulle migliaia di banche dati accessibili on line in tutto il mondo, nonchè sui siti web il cui numero sta crescendo in maniera esponenziale [Za99].

Quest'informazione consiste in pagine di tesi, di memorie, di pubblicazioni scientifiche relative ad un numero imponente di argomenti facilmente accessibili in testo libero on line.

Alcune sono gratuite, anche se la maggioranza di quelle più interessanti sono a pagamento, raccolte attraverso i servizi di qualche infoprovider come EINS (European Information Network Services), Dialog o Lexis/Nexis. Ogni consultazione di questi documenti è fatturata (da pochi centesimi a qualche centinaia di dollari).

I servizi di intelligence (militare e non) sono stati i primi ad interessarsi a strumenti e metodologie in grado di *triturare* il contenuto di queste data banks per estrarne non tanto informazioni puntuali (che si

troverebbero senza grossi problemi) ma, assai più intelligentemente, informazioni di livello superiore, quelle che consistono nelle correlazioni delle informazioni tra di loro. Sia che riguardino luoghi, dichiarazioni, nomi di persone o concetti.

Gli USA cominciarono, attraverso il progetto DARPA (da cui nacque poi Internet), ad interessarsi di questi sistemi fin dagli anni '70. L'obiettivo a quell'epoca era di estrarre le informazioni sensibili da grandi moli di dati testuali in maniera automatica. Qualcuno ricorderà il lavoro svolto dal personaggio interpretato da Robert Redford in "I tre giorni del Condor"...

Inizialmente l'obiettivo era solo di trattare i messaggi (brevi) che arrivavano dall'esterno e di farli arrivare alle persone che più ne erano interessate (attraverso tecniche di *filtering* e *routing*). Ma poi l'interesse si spostò verso sistemi che potevano andare da soli a cercare le informazioni su databases esterni.

Il prodotto *Topic* fu uno dei primi risultati degli sforzi di quest'epoca, e s'impose rapidamente come uno degli standard internazionali. Da allora molti altri prodotti sono stati sviluppati ed utilizzati, prima in campo militare e poi civile, dando nascita alle discipline della Business e Competitive Intelligence.

Ricordiamo prodotti quali Taiga, Noemic, Madicia...[Gu95].

Con l'esplosione del fenomeno di Internet nella seconda metà degli anni '90, si è aperto un nuovo fronte per gli operatori di Intelligence.

Infatti non tutti gli internauti sono coscienti del fatto che, anche se due amici che si scambiano emails abitano nello stesso palazzo, è sufficiente che uno di loro utilizzi come server di posta elettronica quello di un'azienda localizzata in un'altra nazione (Compuserve, Freesurf, Yahoo...) perchè il loro carteggio sia disponibile all'analisi linguistica avanzata dei tools di text mining dei responsabili della sicurezza di quel paese.

Con batterie di filtri informatici e di analizzatori semantici, diviene possibile creare in maniera automatica, da una raccolta di documenti immensa ed eterogenea, costituita da emails, siti web e banche dati, nuove informazioni come, ad esempio, le tendenze di opinione di un certo gruppo di individui.

Ed è inutile sottolineare la potenza di intelligence che l'utilizzazione di questi sistemi di analisi, accoppiata alla capacità di superare firewalls e sistemi di crittografia, potrebbe assicurare ad una organizzazione interessata ad accedere alle informazioni più riservate presenti in rete.

Per ora si conoscono bene le applicazioni che di queste tecnologie stano facendo aziende come IBM, Unilever, Aerospaziale, Michelin, Pfizer, TIM... Fondamentalmente queste aziende, grazie a questi sistemi, conoscono esattamente i trend evolutivi di qualunque tecnologia o strategia senza essere obbligati a fare un noioso lavoro di selezione manuale e di lettura [Wi97].

Alla fine degli anni '90 sono nati prodotti come Text Knowledge Miner, di IBM, Semio Map, di Semio, DR-LINK, di Textwise.

Prodotti di nuova generazione, che cercano di implementare l'analisi semantica, ovvero la capacità di "capire" il significato dei documenti, risolvendo i problemi nascenti dalle ambiguità del linguaggio, delle sue forme retoriche, delle antonomasie, metafore, anafore e paratassi.

Secondo Pascal Coupet, sviluppatore di uno dei prodotti più originali della fine degli anni '90 (TKM, di IBM) ed ora fondatore e CTO di Temis (Text Mining Solutions), siamo in presenza di un salto tecnologico: "L'informatica, dopo aver automatizzato negli anni '70 ed '80 il settore secondario, ovvero le attività dedicate alla produzione, negli anni '90 il terziario, ovvero i servizi, come banche ed assicurazioni, si è lanciata con forza con l'avvio del nuovo millennio nel terziario avanzato, o quaternario: l'automazione, attraverso il text mining, di tutto ciò che riguarda il marketing, la consulenza, l'entertainment. Ad esempio, la nostra società, Temis, ha realizzato un sistema di supporto all'individuazione delle opportunità di business in tutto il mondo per le piccole e medie imprese calabresi nei settori dell'artigianato e del turismo. Opportunità che sono alla portata di tutti, ben descritte in migliaia di pagine su internet. Il problema era identificarle nel mare magnum di milioni, miliardi di pagine disponibili su siti web e all'interno di circa cinquemila banche dati online. Attraverso la nostra soluzione per il text mining semantico online, Online Miner, l'abbiamo reso loro possibile in collaborazione con Telcal, consorzio per lo sviluppo della Regione Calabria, nel corso del 2001".

PARTE TERZA: TEXT MINING – UN’OVERVIEW TECNICA

11. Introduzione

Le ragioni dell’attuale successo del text mining (l’accoppiamento della tecnologia della lingua con gli algoritmi del data mining), sono da ricercarsi:

- 1) nei recenti progressi delle tecniche di NLP (Natural Language Processing) e nella loro formalizzazione matematica,
- 2) nella disponibilità di applicazioni complesse e di potenza elaborativa attraverso gli ASPs (Application Services Providers),
- 3) nell’attenzione corrente di accademici, multinazionali del SW, produttori di motori di ricerca verso tecniche di gestione della lingua, che ci fanno prevedere un forte sviluppo di questa tecnologia nei prossimi mesi.

Text mining è una forma particolare di data mining dove i dati, consistenti in testi liberi, sono destrutturati [BST00]. A causa della presenza del linguaggio naturale, lo step di preparazione dati è più lungo del solito e richiede una fase di preprocessing linguistico [BRZ99], [GG98] per sciogliere, almeno parzialmente, le ambiguità legate alla comprensione del significato. In questo senso il text mining estende il concetto di *content analysis* [We90], che non prevedeva la possibilità di trattare il testo prima di analizzarlo, nella direzione mostrata da [Man97].

L’obiettivo rimane, comunque, lo stesso: scoprire la conoscenza nascosta nei dati, in questo caso documenti, senza definire in anticipo l’argomento della ricerca o definendolo in maniera generica.

La soluzione del text mining, pur avvantaggiandosi delle tecnologie dei motori di ricerca, differisce da questi ultimi che si ripromettono solo di offrire modi più efficienti di recuperare documenti una volta che sia stato definito l’argomento oggetto della ricerca [CDAR98].

12. Fase di preprocessing linguistico

Utilizzando dizionari elettronici, taggers sintattici, motori di lemmatizzazione, il text mining:

- risolve le principali ambiguità legate alla lingua; in questa fase, ad esempio, le parole:
 - *record*, nella frase inglese ‘we record the record’,
 - *couvent*, nella frase francese ‘les poules du couvent couvent’,
 - *pesca*, nella frase italiana ‘pesca la pesca’,sono riconosciute nel loro significato;
- lemmatizza parole/espressioni (“International Business Machines” è trasformato in “IBM”),
- indicizza automaticamente i documenti, associando loro i concetti chiave ivi contenuti.

13. Fase di scoperta di regole

Ove i dati sono trattati seguendo l’approccio classico del data mining [Za97], individuando connessioni/legami/somiglianze tra i diversi temi e ove possono essere individuati i concetti giudicati interessanti.

Ad esempio: *partnerships* siglate (con i nomi delle aziende che le hanno siglate) anche se la parola *partnership* o il nome dell’azienda non appaiono in maniera esplicita nei testi.

14. Le Fonti dei Dati

- Web Data

Internet sta diventando il principale "media" attraverso cui è possibile ottenere documenti, dati ed informazioni. I siti web liberamente raggiungibili via Internet sono una delle fonti principali della documentazione da analizzare.

Questa grande quantità di informazioni, normalmente gratuita e non certificata, è di scarso utilizzo senza l’attuazione di opportune politiche di “filtro”.

- Banche dati online

Le banche dati online costituiscono collezioni di informazioni specializzate, generalmente accessibili via Internet tramite abbonamento. Esempi tipici di queste banche dati sono quelle dedicate alle pubblicazioni, ai brevetti o agli articoli scientifici (di chimica, fisica o matematica) rese disponibili in modo diretto o attraverso information broker.

- Sorgenti informative private

Una banca dati privata di documenti elettronici (costruita negli anni) può essere resa disponibile ed essere opportunamente usata insieme alle altre sorgenti informative. Il formato ed i contenuti dei documenti di una banca dati privata sono generalmente completamente differenti da quelli dei documenti ottenuti attraverso le banche dati online.

- Emails

Le emails sono la forma più ricca dal punto di vista informativo e più semplice da analizzare. E' il mezzo attraverso cui le persone comunicano all'interno ed all'esterno di aziende ed organizzazioni. Possono essere analizzate sia le emails interne ad una organizzazione sia quelle ricevute dall'esterno od inviate all'esterno dell'organizzazione.

Naturalmente queste attività dovranno essere effettuate nel rispetto delle leggi sulla data privacy.

- Opinion surveys

Spesso le opinion surveys sono analizzate con cura nella parte codificata, dove è prevista la risposta: SI, NO, o numerica. Sono invece analizzate in maniera superficiale nella parte testuale, ove si raccolgono le risposte in testo libero alle domande aperte.

- Newsgroups, Chatlines, Mailing Lists

Nuovissima fonte di informazione sono i newsgroups, le chatlines, le mailing lists riguardanti i temi più disparati, dai consumi alla politica. Il problema con questo tipo di informazione è che l'informazione pertinente è all'interno di frasi e/o affermazioni di scarsa importanza, espresse con linguaggio spesso gergale. Grazie al text mining queste affermazioni/opinioni possono essere analizzate e filtrate al fine di conoscere quali sono le opinioni di chi scrive.

15. Le tecniche.

Il Text Mining è una tecnologia complessa e recente e non possiamo aspettarci che tutti i tools disponibili abbiano tutte le possibili funzionalità. Comunque è abbastanza corretto presentare quali sono le tecniche principali che si trovano disponibili sul mercato, anche se non tutti i prodotti disponibili le presentano.

Classificazione. Questa funzione permette di assegnare testi a categorie predefinite, basandosi sul loro contenuto (ad esempio, per permettere che una lettera sia rapidamente assegnata all'ufficio corretto anche se il nome di quest'ufficio non appare formalmente sulla lettera).

Clustering. Permette di raggruppare i documenti in "clusters" basati sul loro contenuto, senza averlo precedentemente definito (ad esempio, migliaia di emails possono essere divise in 20 gruppi, basandosi sul loro contenuto: richiesta di informazioni, reclami, errori, abbonamenti ...).

Estrazione di Termini. Per estrarre la parola o il concetto, in base alla sua categoria (ad esempio, nomi di aziende).

Estrazione di Relazioni. Per estrarre la relazione contenuta nel testo (Bill Gates-President-Microsoft, Rome-Capital-Italy).

Estrazione di Segnali Deboli. Per estrarre espressioni (ad esempio, la frase "Agnelli sposa la General Motors" è riconosciuta come "partnership tra FIAT e GM").

Identificazione di Sigle. Permette il riconoscimento di sigle (ad esempio, NEC è riconosciuto come Nippon Electrical Company, e non come parte della frase "nec plus ultra").

Riconoscimento di Sinonimi. Permette di riconoscere “auto” e “vettura” come sinonimi.

Riconoscimento di Polisemi. Permette di riconoscere che la parola “pesca”, nella frase “pesca la pesca”, ha un doppio significato.

Linguaggi Specifici. Permette il riconoscimento immediato di abbreviazioni, sinonimi e polisemi, individuando la lingua specializzata relativa al documento in oggetto (settore bancario, assicurativo, competitive intelligence, ...), utilizzando thesauri specializzati e dizionari.

PARTE QUARTA: APPLICAZIONI ED ESEMPI

Quali sono nella pratica, le applicazioni nell’intelligence militare e quali nell’intelligence civile (o aziendale)?

Dal punto di vista tecnologico le differenze sono minime o, ancor meglio, non ci sono affatto. Possiamo dire che la differenza sta nel come si accede ai dati. Se nel mondo civile devono essere tenute presenti le esigenze della data privacy nonché le leggi correnti dello stato in cui si opera, dal punto di vista militare queste tematiche sono meno vincolanti in quanto, per sua natura, l’operazione militare si svolge in situazione di emergenza, in cui i vincoli di rispetto delle leggi sono, quantomeno, d’altro tipo.

Ma vediamo alcune delle applicazioni, alla luce di quanto scritto in precedenza.

Si è deciso di presentare alcune delle applicazioni in ambito militare ed un caso concreto di applicazione al settore civile, in particolare al servizio di competitive intelligence in IBM, azienda che l’utilizza regolarmente nelle sue attività globali di market intelligence.

16. Applicazioni militari

- Anti-terrorismo tecnologico.

Sarà pur vero che ormai le armi nucleari, chimiche e biologiche sono accessibili a chiunque, ma in questa frase è presente, ovviamente, una semplificazione. Se un coltello, o una pistola, sono veramente reperibili da chiunque, la capacità di produrre o di acquisire, trasportare e diffondere materiale radiattivo o gas nervini o, ancora, il bacillo della peste polmonare, è immaginabile sia a disposizione di molti ma non proprio di tutti. Dovrà essere sempre coinvolto uno specialista di queste sostanze, che avrà studiato in qualche scuola od università, che forse avrà compiuto ricerche e scritto articoli o libri e che probabilmente, anche in anni precedenti, avrà partecipato a convegni e congressi su questi temi.

Questo significa che avrà lasciato, come altre centinaia o migliaia di persone come lui, delle tracce elettroniche dei suoi interessi, dei suoi legami, dei suoi collegamenti con movimenti religiosi, gruppi politici, organizzazioni criminali. Ad esempio, se stiamo monitorando i movimenti dei cittadini che lavorano all’estero (cosa piuttosto elementare da farsi) di uno stato particolarmente aggressivo nell’area mediorientale, possiamo scoprire che una ventina di loro stanno concentrandosi in un edificio ben specifico di questo paese mediorientale. E se le aree di specializzazione di questi individui sono aree abbastanza innocue prese isolatamente, ma prese insieme permettono di ricostruire la catena di produzione di un gas nervino, allora avremo ragione di dubitare delle affermazioni dei leaders di quel paese quando dichiarano che, in quell’edificio, si studiano nuovi processi per la produzione di latte in polvere.

Le tecniche di text mining che permettono l’individuazione di questi “collegi invisibili” di specialisti sono ben conosciute ed utilizzate regolarmente nel mondo industriale per permettere l’individuazione delle strategie di ricerca dei competitors e prevederne la produzione futura [Za98].

- Analisi delle Rivendicazioni.

E’ ormai abbastanza frequente che le uniche tracce di attentati terroristici altro non siano che le relative lettere di rivendicazione, eventualmente trasmesse via rete elettronica. In tal caso neppure il supporto cartaceo potrà essere utilizzato nelle analisi di polizia giudiziaria. Ma una traccia interessante risulta essere

in tal caso lo stile con cui queste lettere sono scritte, i concetti espressi ed i loro collegamenti, in definitiva la struttura dello scritto.

E poichè spesso si sospetta da quali gruppi politici questi attentati provengono, un confronto automatico tra tale scritto e la produzione tipica di, ipotizziamo, 300 di questi gruppi ci permetterebbe di individuare il probabile gruppo di provenienza dell'autore della lettera. E' con questa tipo di tecnologia che alcuni famosi attentatori (quello di Unabomber è il caso più famoso) furono descritti con precisione prima ancora di essere poi veramente individuati.

- Info Spam.

Nello scenario di una guerra informatica una delle armi più conosciute è l'Info Spam (intasamento delle caselle postali elettroniche di messaggi con lo scopo di impedire l'utilizzo della rete informatica da parte del bersaglio di questo attacco). Fu una delle poche azioni che i Serbi riuscirono a portare a termine contro gli USA nel corso della guerra nei Balcani del 1999.

E' anche uno degli effetti che si propongono alcuni tipi di virus, che mettono così in crisi caselle postali, utenti, servers, aziende intere...

Il text mining è in grado di filtrare i messaggi in arrivo, riconoscendo quelli sospetti, senza farli arrivare alla casella.

- Competitive Intelligence.

I servizi di intelligence militare, specialmente negli ultimi anni, hanno spesso supportato le aziende del proprio paese raccogliendo informazioni sulle aziende loro concorrenti (anche se di paesi della medesima alleanza) [Ad94]. Operando sulle cosiddette *fonti aperte* (agenzie stampe, riviste, siti web, rapporti) o su dati (sia testuali che vocali) intercettati il text mining permette di evidenziare quali sono le tendenze strategiche di aziende singole o di complessi industriali o, addirittura, della politica economica di un paese, al di là di quello che viene dichiarato pubblicamente [Ba94]. Queste tecniche sono ormai di dominio pubblico anche nel settore civile [Za00], [Za01].

- Individuazione di attività di lobbying.

Sia che tale attività sia legittima che illegittima, il text mining permette di evidenziare quali giornalisti, quali giornali, quali gruppi di media, quali uomini politici, quali criminali sono, nei fatti, alleati. Se tali alleanze si presentano un numero di volte superiore a quello che statisticamente ci si aspetta può essere avviata un'attività di verifica di eventuali collusioni.

- Controllo di settori specifici.

Una dei compiti dei servizi di intelligence è quello di monitorare varia documentazione per individuare concetti che possano rivestire interesse, avendo già definito in precedenza quali sono tali concetti. Il text mining potrà individuare tali concetti all'interno di grandi volumi di documenti di fonte anche eterogenea, ed evidenziarli, contarli, riassumerli.

- Proliferazione.

Quanto detto in precedenza per il terrorismo tecnologico può essere a maggior ragione affermato per quanto riguarda il controllo della produzione industriale diretta agli armamenti.

- Controllo reticoli sociali

Molte comunicazioni al giorno d'oggi avvengono via Internet, sotto forma di email, mailing list, chat lines, forums, newsgroups, bollettin boards. Il text mining permette di analizzare questa importante messe di informazioni estraendo fatti, citazioni, sentimenti... permettendo l'identificazione di collegamenti nascosti o individuando trends di comportamenti di particolare interesse [Za03].

- Altro.

Qualunque oggetto dell'attività classica di intelligence, se opportunamente disegnato, può beneficiare delle tecniche di text mining che hanno l'obiettivo non di sostituire l'attività dell'analista quanto di automatizzarne l'aspetto più ripetitivo e di minor valore aggiunto.

AUTORE

Laureato in Ingegneria Nucleare all'Università di Bologna, specializzato in matematica applicata a problemi di intelligence all'Università di Paris VI e all'Università di Modena, ove tuttora insegna.

Iniziò la sua attività come Ufficiale dei Carabinieri, al Centro Carabinieri Investigazioni Scientifiche (CCIS) di Roma. Dopo un periodo come consulente tecnico per diverse organizzazioni statali, entrò in IBM dove e' rimasto sino al 2001 e dove, dopo alcuni incarichi in USA (San Jose, CA) e Francia (Paris), ha avuto la responsabilita' della funzione di Market Intelligence nell'area Sud Europa.

Ha coordinato le attività del Bologna KDD Center, basato presso Cineca (consorzio interuniversitario per il calcolo automatico), ha tenuto corsi e conferenze all'Università Bocconi di Milano; all'Università Federale di Rio de Janeiro (UFRJ); alle Scuole di Guerra ed Accademie Militari Italiana, Francese e Brasiliana, alle Scuole dell'Arma dei Carabinieri; a workshop internazionali. Coordina progetti relativi a temi di Informazione Online ed Intelligence sia in Europa che in Nord e Sud America, ed e' attualmente consulente della Commissione Europea su tematiche di business intelligence.

Membro dell'IAFE (International Association of Financial Engineers), dell'European Advisory Board di SCIP (Society of Competitive Intelligence Professionals) e coordinatore del chapter SCIP di Bologna, dell'Associazione Internazionale di Polizia e dell'Associazione Nazionale dei Carabinieri, dell'Associazione Italiana per l'Intelligenza Artificiale, del comitato scientifico della rivista Scienza e Business.

E' autore/coautore di diverse pubblicazioni su argomenti di intelligence elettronica.

BIBLIOGRAFIA

- [Ad94] – J.Adams, 1994 – New Spies - Ed.Pimlico, London
- [Ba94] – W.D.Barndt, 1994 – User-directed Competitive Intelligence – Ed.Quorum Books, Westport, CT
- [BKC95] - Bologna KDD Center - <http://open.cineca.it/datamining>
- [BRZ99] – R.Baeza-Yates, B.Ribeiro-Neto, N.Ziviani, 1999 – Text Operations, in “R.Baeza-Yates, B.Ribeiro-Neto: Modern Information Retrieval” – pp.163-190 – Addison Wesley
- [BST00] – A.Berson, S.Smith, K.Thearling, 2000 – Building Data Mining Applications for CRM – pp.457-484 – McGrawHill
- [CDAR98] – S.Chakrabarty, B.Dom, R.Agrawal, P.Raghavan, 1998 – Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies – The VLDB Journal (1998) 7:163-178
- [Co97] – Leonard A.Cole, 1997 – Lo spettro delle armi biologiche – Le Scienze, n.342, febbraio 1997
- [DGS99] - Doerre, Gerstl, Seiffert, 1999 - Text Mining: Finding Nuggets in Mountains of Textual Data - KDD99 Proceedings – ACM
- [Fa99] – L.Fahey, 1999 – Competitors – Ed.John Wiley
- [FC95] – FC, 1995 - Industrial Society and its Future (The Unabomber Manifesto) – Jolly Roger Press
- [F100] – B.Flynt – Threat Kingdom – Military Review, July/August 2000
- [GG98] - C.Gouffas, J.M.Granier, 1998 - The Death of Lady D - A Semiotic and Psycho-sociological Understanding of the Funeral Eulogy - ESOMAR Seminar on the Internet and Market Research
- [GG96] - C.Gouffas, J.M.Granier, 1996 - Les mots de l'Entreprise: Analyse Textuelle Automatique et Semiotique - JADT Rome 96 Seminar
- [GMS95] – R.Godson, E.R.May, G.Schmitt,1995 - US Intelligence at the Crossroads – Brassey's, Washington
- [Gr97] - C.H.Gray, 1997 - Postmodern War – The Guilford Press, London
- [Gu95] – J.Guisnel, 1995 – Guerres dans le cyberspace – Paris, Editions la Decouverte
- [HL00] – S.Hayward, A.Linden, 2000 – Gartner Group RAS Services – 6/6/2000
- [HQD91] - C.Huot, L.Quoniam, H.Dou, 1991 - A new Method for Analyzing Downloaded Data for Strategic Decision - Scientometrics, Vol.22-No.2
- [Hu96] - C.Huot, 1996 - IBM Technology Watch - IBM Internal Report
- [La96] – W.Laqueur, 1996 – Postmodern Terrorism – Foreign Affairs, Sept./Oct. 1996

- [Ma01] – L.Mainoldi, 2001 – Oltre Echelon: dove va lo spionaggio elettronico – I signori della rete - Suppl.n.1/2001 di Limes
- [Ma91] - J.F.Marcotorchino, 1991 - L'Analyse Factorielle Relationelle (partie I et II) - Etude du CEMAP IBM France - N°MAP-003
- [Ma99] – R.Mattison, 1999 - Web Warehousing and Knowledge Management, pp.337-354 - McGrawHill
- [Man97] – I.Mani et al., 1997 – Towards Content-Based Browsing of Broadcast News Video – in Intelligent Multimedia Information Retrieval - Ed.AAAI Press
- [Ro96] – Jeffrey Robinson, 1996 – The Laundrymen – Arcade Publishing, New York
- [Ro96b] - D.Rouach, 1996 - La veille technologique et l'intelligence économique – Ed.PUF, Paris
- [St01] – A.Stone, 2001 – Troops armed with clicks, keystrokes – USA Today, June 19, 2001
- [Su97] – R.Summers, 1997 – Secure Computing: Threats and Safeguards – Ed.McGrawHill
- [Te00] – <http://www.temis-group.com>
- [Th97] – The Economist, 1997 – The Future of Warfare – March, 8th 1997
- [TT93] – A. and H.Toffler, 1993 – War and Antiwar - Warner Books
- [We90] – R.P.Weber, 1990 – Basic Content Analysis – SAGE Publications
- [Wi97] - I.Winkler,1997 - Corporate Espionage – Prima Publishing
- [YL99] – E.Younger, A.Linden, 1999 – CIO Update:Data Mining Applications of the next Decade – 7/7/1999, Gartner Group Article
- [Za86] – A.Zanasi, 1986 - Computer Crimes – N.6, Rassegna Arma Carabinieri
- [Za87] – A.Zanasi, 1987 – Identificazione di armi nascoste – N.4, Rassegna Arma Carabinieri
- [Za97] - A.Zanasi, 1997 - Discovering Data Mining - Prentice Hall
- [Za98] - A.Zanasi, 1998 - Competitive Intelligence Thru Data Mining Public Sources - Competitive Intelligence Review - Vol.9(1) - John Wiley & Sons, Inc.
- [Za99] – A.Zanasi, 1999 – in Competitive Technology Intelligence – John Wiley & Sons, Inc.
- [Za00] – A.Zanasi, 2000 – Web Mining through the Online Miner - Data Mining 2000 Proceedings – Wessex Institute of Technology
- [Za01] – A.Zanasi, 2001 – Text Mining: the new CI Frontier – *VSSST 2001* – Ed.IRIT
- [Za02] – A.Zanasi, 2002 – Data Mining III – Wessex Institute of Technology
- [Za02b] – A.Zanasi, 2000–Applications of High Performance Computing – Wessex Institute of Technology
- [Za03] – A.Zanasi, 2003 - Email, chatlines, newsgroups: a continuous opinion surveys source thanks to text mining application – *Excellence 2003 in Int'l Research* - Ed.Esomar